PCT/US05/02379

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 15/11, 15/10, C07K 14/47, C12P 21/00, C12Q 1/68, C07K 16/18, G06F 17/30, 17/50

A2

(11) International Publication Number: **WO 99/53051**

(43) International Publication Date: 21 October 1999 (21.10.99)

(21) International Application Number: PCT/IB99/00712

(22) International Filing Date: 9 April 1999 (09.04.99)

(30) Priority Data:

09/057,719	9 April 1998 (09.04.98)	US
09/069,047	28 April 1998 (28.04.98)	US

(81) Designated States: AU, CA, JP, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(71) Applicant *(for all designated States except US)*: GENSET [FR/FR]; 24, rue Royale, F–75008 Paris (FR).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: DUMAS MILNE EDWARDS, Jean–Baptiste [FR/FR]; 8, rue Grégoire–de–Tours, F–75006 Paris (FR). DUCLERT, Aymeric [FR/FR]; 6 ter, rue Victorine, F–94100 Saint–Maur (FR). GIORDANO, Jean–Yves [FR/FR]; 12, rue Duhesme, F–75018 Paris (FR).

(74) Agents: MARTIN, Jean–Jacques et al.; Cabinet Regimbeau, 26, avenue Kléber, F–75116 Paris (FR).

(54) Title: 5' ESTS AND ENCODED HUMAN PROTEINS

(57) Abstract

The sequences of 5' ESTs derived from mRNAs encoding secreted proteins are disclosed. The 5' ESTs may be to obtain cDNAs and genomic DNAs corresponding to the 5' ESTs. The 5' ESTs may also be used in diagnostic, forensic, gene therapy, and chromosome mapping procedures. Upstream regulatory sequences may also be otained using the 5' ESTs. The 5' ESTs may also be used to design expression vectors and secretion vectors.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

5' ESTS AND ENCODED HUMAN PROTEINS

Background of the Invention

The estimated 50,000-100,000 genes scattered along the human chromosomes offer tremendous promise for the understanding, diagnosis, and treatment of human diseases. In addition, probes capable of specifically hybridizing to loci distributed throughout the human genome find applications in the construction of high resolution chromosome maps and in the identification of individuals.

In the past, the characterization of even a single human gene was a painstaking process, requiring years of effort. Recent developments in the areas of cloning vectors, DNA sequencing, and computer technology have merged to greatly accelerate the rate at which human genes can be isolated, sequenced, mapped, and characterized.

Currently, two different approaches are being pursued for identifying and characterizing the genes distributed along the human genome. In one approach, large fragments of genomic DNA are isolated, cloned, and sequenced. Potential open reading frames in these genomic sequences are identified using bioinformatics software. However, this approach entails sequencing large stretches of human DNA which do not encode proteins in order to find the protein encoding sequences scattered throughout the genome. In addition to requiring extensive sequencing, the bioinformatics software may mischaracterize the genomic sequences obtained, *i.e.*, labeling non-coding DNA as coding DNA and vice versa.

An alternative approach takes a more direct route to identifying and characterizing human genes. In this approach, complementary DNAs (cDNAs) are synthesized from isolated messenger RNAs (mRNAs) which encode human proteins. Using this approach, sequencing is only performed on DNA which is derived from protein coding portions of the genome. Often, only short stretches of the cDNAs are sequenced to obtain sequences called expressed sequence tags (ESTs). The ESTs may then be used to isolate or purify extended cDNAs which include sequences adjacent to the EST sequences. The extended cDNAs may contain all of the sequence of the EST which was used to obtain them or only a portion of the sequence of the EST which was used to obtain them. In addition, the extended cDNAs may contain the full coding sequence of the gene from which the EST was derived or, alternatively, the extended cDNAs may include portions of the coding sequence of the gene from which the EST was derived. It will be appreciated that there may be several extended cDNAs which include the EST sequence as a result of alternate splicing or the activity of alternative promoters. Alternatively, ESTs having partially overlapping sequences may be identified and contigs comprising the consensus sequences of the overlapping ESTs may be identified.

In the past, these short EST sequences were often obtained from oligo-dT primed cDNA libraries. Accordingly, they mainly corresponded to the 3' untranslated region of the mRNA. In part, the prevalence of EST sequences derived from the 3' end of the mRNA is a result of the fact that typical

techniques for obtaining cDNAs, are not well suited for isolating cDNA sequences derived from the 5' ends of mRNAs (Adams *et al., Nature* 377:3-174, 1996, Hillier *et al., Genome Res.* 6:807-828, 1996).

In addition, in those reported instances where longer cDNA sequences have been obtained, the reported sequences typically correspond to coding sequences and do not include the full 5' untranslated region (5'UTR) of the mRNA from which the cDNA is derived. Indeed, 5'UTRs have been shown to affect either the stability or translation of mRNAs. Thus, regulation of gene expression may be achieved through the use of alternative 5'UTRs as shown, for instance, for the translation of the tissue inhibitor of metalloprotease mRNA in mitogenically activated cells (Waterhouse *et al, J Biol Chem.* **265**:5585-9. 1990). Furthermore, modification of 5'UTR through mutation, insertion or translocation events may even be implied in pathogenesis. For instance, the fragile X syndrome, the most common cause of inherited mental retardation, is partly due to an insertion of multiple CGG trinucleotides in the 5'UTR of the fragile X mRNA resulting in the inhibition of protein synthesis via ribosome stalling (Feng *et al, Science* **268**:731-4, 1995). An aberrant mutation in regions of the 5'UTR known to inhibit translation of the proto-oncogene *c-myc* was shown to result in upregulation of c-myc protein levels in cells derived from patients with multiple myelomas (Willis *et al, Curr Top Microbiol Immunol* **224**:269-76, 1997). In addition, the use of oligo-dT primed cDNA libraries does not allow the isolation of complete 5'UTRs since such incomplete sequences obtained by this process may not include the first exon of the mRNA, particularly in situations where the first exon is short. Furthermore, they may not include some exons, often short ones, which are located upstream of splicing sites. Thus, there is a need to obtain sequences derived from the 5' ends of mRNAs.

While many sequences derived from human chromosomes have practical applications, approaches based on the identification and characterization of those chromosomal sequences which encode a protein product are particularly relevant to diagnostic and therapeutic uses. In some instances, the sequences used in such therapeutic or diagnostic techniques may be sequences which encode proteins which are secreted from the cell in which they are synthesized. Those sequences encoding secreted proteins as well as the secreted proteins themselves, are particularly valuable as potential therapeutic agents. Such proteins are often involved in cell to cell communication and may be responsible for producing a clinically relevant response in their target cells. In fact, several secretory proteins, including tissue plasminogen activator, G-CSF, GM-CSF, erythropoietin, human growth hormone, insulin, interferon-α, interferon-β, interferon-γ, and interleukin-2, are currently in clinical use. These proteins are used to treat a wide range of conditions, including acute myocardial infarction, acute ischemic stroke, anemia, diabetes, growth hormone deficiency, hepatitis, kidney carcinoma, chemotherapy-induced neutropenia and multiple sclerosis. For these reasons, extended cDNAs encoding secreted proteins or portions thereof represent a valuable source of therapeutic agents. Thus, there is a need for the identification and characterization of secreted proteins and the nucleic acids encoding them.

In addition to being therapeutically useful themselves, secretory proteins include short peptides, called signal peptides, at their amino termini which direct their secretion. These signal peptides are

encoded by the signal sequences located at the 5' ends of the coding sequences of genes encoding secreted proteins. These signal peptides can be used to direct the extracellular secretion of any protein to which they are operably linked. In addition, portions of the signal peptides called membrane-translocating sequences, may also be used to direct the intracellular import of a peptide or protein of interest. This may prove beneficial in gene therapy strategies in which it is desired to deliver a particular gene product to cells other than the cells in which it is produced. Signal sequences encoding signal peptides also find application in simplifying protein purification techniques. In such applications, the extracellular secretion of the desired protein greatly facilitates purification by reducing the number of undesired proteins from which the desired protein must be selected. Thus, there exists a need to identify and characterize the 5' portions of the genes for secretory proteins which encode signal peptides.

Sequences coding for non-secreted proteins may also find application as therapeutics or diagnostics. In particular, such sequences may be used to determine whether an individual is likely to express a detectable phenotype, such as a disease, as a consequence of a mutation in the coding sequence of a protein. In instances where the individual is at risk of suffering from a disease or other undesirable phenotype as a result of a mutation in such a coding sequence, the undesirable phenotype may be corrected by introducing a normal coding sequence using gene therapy. Alternatively, if the undesirable phenotype results from overexpression of the protein encoded by the coding sequence, expression of the protein may be reduced using antisense or triple helix based strategies.

The secreted or non-secreted human polypeptides encoded by the coding sequences may also be used as therapeutics by administering them directly to an individual having a condition, such as a disease, resulting from a mutation in the sequence encoding the polypeptide. In such an instance, the condition can be cured or ameliorated by administering the polypeptide to the individual.

In addition, the secreted or non-secreted human polypeptides or portions thereof may be used to generate antibodies useful in determining the tissue type or species of origin of a biological sample. The antibodies may also be used to determine the cellular localization of the secreted or non-secreted human polypeptides or the cellular localization of polypeptides which have been fused to the human polypeptides. In addition, the antibodies may also be used in immunoaffinity chromatography techniques to isolate, purify, or enrich the human polypeptide or a target polypeptide which has been fused to the human polypeptide.

Public information on the number of human genes for which the promoters and upstream regulatory regions have been identified and characterized is quite limited. In part, this may be due to the difficulty of isolating such regulatory sequences. Upstream regulatory sequences such as transcription factor binding sites are typically too short to be utilized as probes for isolating promoters from human genomic libraries. Recently, some approaches have been developed to isolate human promoters. One of them consists of making a CpG island library (Cross *et al.*, *Nature Genetics* 6: 236-244, 1994). The second consists of isolating human genomic DNA sequences containing SpeI binding sites by the use of SpeI binding protein. (Mortlock *et al.*, *Genome Res.* 6:327-335, 1996). Both of these approaches have

their limits due to a lack of specificity and of comprehensiveness. Thus, there exists a need to identify and systematically characterize the 5' portions of the genes.

The present 5' ESTs may be used to efficiently identify and isolate 5'UTRs and upstream regulatory regions which control the location, developmental stage, rate, and quantity of protein synthesis, as well as the stability of the mRNA. Once identified and characterized, these regulatory regions may be utilized in gene therapy or protein purification schemes to obtain the desired amount and locations of protein synthesis or to inhibit, reduce, or prevent the synthesis of undesirable gene products.

In addition, ESTs containing the 5' ends of protein genes may include sequences useful as probes for chromosome mapping and the identification of individuals. Thus, there is a need to identify and characterize the sequences upstream of the 5' coding sequences of genes.

Summary of the Invention

The present invention relates to purified, isolated, or enriched 5' ESTs which include sequences derived from the authentic 5' ends of their corresponding mRNAs. The term "corresponding mRNA" refers to the mRNA which was the template for the cDNA synthesis which produced the 5' EST. These sequences will be referred to hereinafter as "5' ESTs." The present invention also includes purified, isolated or enriched nucleic acids comprising contigs assembled by determining a consensus sequences from a plurality of ESTs containing overlapping sequences. These contigs will be referred to herein as "consensus contigated 5'ESTs."

As used herein, the term "purified" does not require absolute purity; rather, it is intended as a relative definition. Individual 5' EST clones isolated from a cDNA library have been conventionally purified to electrophoretic homogeneity. The sequences obtained from these clones could not be obtained directly either from the library or from total human DNA. The cDNA clones are not naturally occurring as such, but rather are obtained via manipulation of a partially purified naturally occurring substance (messenger RNA). The conversion of mRNA into a cDNA library involves the creation of a synthetic substance (cDNA) and pure individual cDNA clones can be isolated from the synthetic library by clonal selection. Thus, creating a cDNA library from messenger RNA and subsequently isolating individual clones from that library results in an approximately 10^4-10^6 fold purification of the native message. Purification of starting material or natural material to at least one order of magnitude, preferably two or three orders, and more preferably four or five orders of magnitude is expressly contemplated.

As used herein, the term "isolated" requires that the material be removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide present in a living animal is not isolated, but the same polynucleotide, separated from some or all of the coexisting materials in the natural system, is isolated.

As used herein, the term "recombinant" means that the 5' EST is adjacent to "backbone" nucleic acid to which it is not adjacent in its natural environment. Additionally, to be "enriched" the 5' ESTs will

represent 5% or more of the number of nucleic acid inserts in a population of nucleic acid backbone molecules. Backbone molecules according to the present invention include nucleic acids such as expression vectors, self-replicating nucleic acids, viruses, integrating nucleic acids, and other vectors or nucleic acids used to maintain or manipulate a nucleic acid insert of interest. Preferably, the enriched 5' ESTs represent 15% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules. More preferably, the enriched 5' ESTs represent 50% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules. In a highly preferred embodiment, the enriched 5' ESTs represent 90% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules.

"Stringent," "moderate," and "low" hybridization conditions are as defined below.

The term "polypeptide" refers to a polymer of amino acids without regard to the length of the polymer; thus, peptides, oligopeptides, and proteins are included within the definition of polypeptide. This term also does not specify or exclude post-expression modifications of polypeptides, for example, polypeptides which include the covalent attachment of glycosyl groups, acetyl groups, phosphate groups, lipid groups and the like are expressly encompassed by the term polypeptide. Also included within the definition are polypeptides which contain one or more analogs of an amino acid (including, for example, non-naturally occurring amino acids, amino acids which only occur naturally in an unrelated biological system, modified amino acids from mammalian systems etc.), polypeptides with substituted linkages, as well as other modifications known in the art, both naturally occurring and non-naturally occurring.

As used interchangeably herein, the terms "nucleic acids," "oligonucleotides," and "polynucleotides" include RNA, DNA, or RNA/DNA hybrid sequences of more than one nucleotide in either single chain or duplex form. The term "nucleotide" as used herein as an adjective to describe molecules comprising RNA, DNA, or RNA/DNA hybrid sequences of any length in single-stranded or duplex form. The term "nucleotide" is also used herein as a noun to refer to individual nucleotides or varieties of nucleotides, meaning a molecule, or individual unit in a larger nucleic acid molecule, comprising a purine or pyrimidine, a ribose or deoxyribose sugar moiety, and a phosphate group, or phosphodiester linkage in the case of nucleotides within an oligonucleotide or polynucleotide. Although the term "nucleotide" is also used herein to encompass "modified nucleotides" which comprise at least one modifications (a) an alternative linking group, (b) an analogous form of purine, (c) an analogous form of pyrimidine, or (d) an analogous sugar, for examples of analogous linking groups, purine, pyrimidines, and sugars see for example PCT publication No. WO 95/04064. The polynucleotide sequences of the invention may be prepared by any known method, including synthetic, recombinant, *ex vivo* generation, or a combination thereof, as well as utilizing any purification methods known in the art.

The terms "base paired" and "Watson & Crick base paired" are used interchangeably herein to refer to nucleotides which can be hydrogen bonded to one another be virtue of their sequence

identities in a manner like that found in double-helical DNA with thymine or uracil residues linked to adenine residues by two hydrogen bonds and cytosine and guanine residues linked by three hydrogen bonds (See Stryer, L., *Biochemistry*, 4th edition, 1995).

The terms "complementary" or "complement thereof" are used herein to refer to the sequences of polynucleotides which are capable of forming Watson & Crick base pairing with another specified polynucleotide throughout the entirety of the complementary region. For the purpose of the present invention, a first polynucleotide is deemed to be complementary to a second polynucleotide when each base in the first polynucleotide is paired with its complementary base. Complementary bases are, generally, A and T (or A and U), or C and G. "Complement" is used herein as a synonym from "complementary polynucleotide," "complementary nucleic acid" and "complementary nucleotide sequence" . These terms are applied to pairs of polynucleotides based solely upon their sequences and not any particular set of conditions under which the two polynucleotides would actually bind. Preferably, a "complementary" sequence is a sequence which an A at each position where there is a T on the opposite strand, a T at each position where there is an A on the opposite strand, a G at each position where there is a C on the opposite strand and a C at each position where there is a G on the opposite strand.

Thus, 5' ESTs in cDNA libraries in which one or more 5' ESTs make up 5% or more of the number of nucleic acid inserts in the backbone molecules are "enriched recombinant 5' ESTs" as defined herein. Likewise, 5' ESTs in a population of plasmids in which one or more 5' ESTs of the present invention have been inserted such that they represent 5% or more of the number of inserts in the plasmid backbone are "enriched recombinant 5' ESTs" as defined herein. However, 5' ESTs in cDNA libraries in which 5' ESTs constitute less than 5% of the number of nucleic acid inserts in the population of backbone molecules, such as libraries in which backbone molecules having a 5' EST insert are extremely rare, are not "enriched recombinant 5' ESTs."

In some embodiments, the present invention relates to 5' ESTs which are derived from genes encoding secreted proteins. As used herein, a "secreted" protein is one which, when expressed in a suitable host cell, is transported across or through a membrane, including transport as a result of signal peptides in its amino acid sequence. "Secreted" proteins include without limitation proteins secreted wholly (e.g. soluble proteins), or partially (e.g. receptors) from the cell in which they are expressed. "Secreted" proteins also include without limitation proteins which are transported across the membrane of the endoplasmic reticulum.

Such 5' ESTs include nucleic acid sequences, called signal sequences, which encode signal peptides which direct the extracellular secretion of the proteins encoded by the genes from which the 5' ESTs are derived. Generally, the signal peptides are located at the amino termini of secreted proteins.

Secreted proteins are translated by ribosomes associated with the "rough" endoplasmic reticulum. Generally, secreted proteins are co-translationally transferred to the membrane of the endoplasmic reticulum. Association of the ribosome with the endoplasmic reticulum during translation

of secreted proteins is mediated by the signal peptide. The signal peptide is typically cleaved following its co-translational entry into the endoplasmic reticulum. After delivery to the endoplasmic reticulum, secreted proteins may proceed through the Golgi apparatus. In the Golgi apparatus, the proteins may undergo post-translational modification before entering secretory vesicles which transport them across the cell membrane.

The 5' ESTs of the present invention have several important applications. For example, they may be used to obtain and express cDNA clones which include the full protein coding sequences of the corresponding gene products, including the authentic translation start sites derived from the 5' ends of the coding sequences of the mRNAs from which the 5' ESTs are derived. These cDNAs will be referred to hereinafter as "full-length cDNAs." These cDNAs may also include DNA derived from mRNA sequences upstream of the translation start site. The full-length cDNA sequences may be used to express the proteins corresponding to the 5' ESTs. As discussed above, secreted proteins and non-secreted proteins may be therapeutically important. Thus, the proteins expressed from the cDNAs may be useful in treating and controlling a variety of human conditions. The 5' ESTs may also be used to obtain the corresponding genomic DNA. The term "corresponding genomic DNA" refers to the genomic DNA which encodes the mRNA from which the 5' EST was derived.

Alternatively, the 5' ESTs may be used to obtain and express extended cDNAs encoding portions of the protein. In the case of secreted proteins, the portions may comprise the signal peptides of the secreted proteins or the mature proteins generated when the signal peptide is cleaved off.

The present invention includes isolated, purified, or enriched "EST-related nucleic acids." The terms "isolated," "purified" or "enriched" have the meanings provided above. As used herein, the term "EST-related nucleic acids" means the nucleic acids of SEQ ID NOs. 24-811 and 1600-1622, extended cDNAs obtainable using the nucleic acids of SEQ ID NOs. 24-811 and 1600-1622, full-length cDNAs obtainable using the nucleic acids of SEQ ID NOs. 24-811 and 1600-1622 or genomic DNAs obtainable using the nucleic acids of SEQ ID NOs. 24-811 and 1600-1622. The present invention also includes the sequences complementary to the EST-related nucleic acids.

The present invention also includes isolated, purified, or enriched "fragments of EST-related nucleic acids." The terms "isolated," "purified" and "enriched" have the meanings described above. As used herein the term "fragments of EST-related nucleic acids" means fragments comprising at least 10, 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, 50, 75, 100, 200, 300, 500, or 1000 consecutive nucleotides of the EST-related nucleic acids to the extent that fragments of these lengths are consistent with the lengths of the particular EST-related nucleic acids being referenced. In particular, fragments of EST-related nucleic acids refer to "polynucleotides described in Table II," "polynucleotides described in Table III," and "polynucleotides described in Table IV." The present invention also includes the sequences complementary to the fragments of the EST-related nucleic acids.

The present invention also includes isolated, purified, or enriched "positional segments of EST-related nucleic acids." As used herein, the term "positional segments of EST-related nucleic acids"

includes segments comprising nucleotides 1-25, 26-50, 51-75, 76-100, 101-125, 126-150, 151-175, 176-200, 201-225, 226-250, 251-300, 301-325, 326-350, 351-375, 376-400, 401-425, 426-450, 451-475, 476-500, 501-525, 526-550, 551-575, 576-600 and 601-the terminal nucleotide of the EST-related nucleic acids to the extent that such nucleotide positions are consistent with the lengths of the particular EST-related nucleic acids being referenced. The term "positional segments of EST-related nucleic acids also includes segments comprising nucleotides 1-50, 51-100, 101-150, 151-200, 201-250, 251-300, 301-350, 351-400, 401-450, 450-500, 501-550, 551-600 or 601-the terminal nucleotide of the EST-related nucleic acids to the extent that such nucleotide positions are consistent with the lengths of the particular EST-related nucleic acids being referenced. The term "positional segments of EST-related nucleic acids" also includes segments comprising nucleotides 1-100, 101-200, 201-300, 301-400, 501-500, 500-600, or 601-the terminal nucleotide of the EST-related nucleic acids to the extent that such nucleotide positions are consistent with the lengths of the particular EST-related nucleic acids being referenced. In addition, the term "positional segments of EST-related nucleic acids" includes segments comprising nucleotides 1-200, 201-400, 400-600, or 601-the terminal nucleotide of the EST-related nucleic acids to the extent that such nucleotide positions are consistent with the lengths of the particular EST-related nucleic acids being referenced. The present invention also includes the sequences complementary to the positional segments of EST-related nucleic acids.

The present invention also includes isolated, purified, or enriched "fragments of positional segments of EST-related nucleic acids." As used herein, the term "fragments of positional segments of EST-related nucleic acids" refers to fragments comprising at least 10, 15, 18, 20, 23, 25, 28, 30, 35, 40, 50, 75, 100, 150, or 200 consecutive nucleotides of the positional segments of EST-related nucleic acids. The present invention also includes the sequences complementary to the fragments of positional segments of EST-related nucleic acids.

The present invention also includes isolated or purified "EST-related polypeptides." As used herein, the term "EST-related polypeptides" means the polypeptides encoded by the EST-related nucleic acids, including the polypeptides of SEQ ID NOs. 812-1599.

The present invention also includes isolated or purified "fragments of EST-related polypeptides." As used herein, the term "fragments of EST-related polypeptides" means fragments comprising at least 5, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of an EST-related polypeptide to the extent that fragments of these lengths are consistent with the lengths of the particular EST-related polypeptides being referenced. In particular, fragments of EST-related polypepides refer to polypeptides encoded by "polynucleotides described in Table II," "polynucleotides described in Table III," and "polynucleotides described in Table IV."

The present invention also includes isolated or purified "positional segments of EST-related polypeptides." As used herein, the term "positional segments of EST-related polypeptides" includes polypeptides comprising amino acid residues 1-25, 26-50, 51-75, 76-100, 101-125, 126-150, 151-175, 176-200, or 201-the C-terminal amino acid of the EST-related polypeptides to the extent that such amino

nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids may be useful in treating or controlling a variety of human conditions.

The EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids may be used in forensic procedures to identify individuals or in diagnostic procedures to identify individuals having genetic diseases resulting from abnormal gene expression. In addition, the EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids are useful for constructing a high resolution map of the human chromosomes.

The present invention also relates to secretion vectors capable of directing the secretion of a protein of interest. Such vectors may be used in gene therapy strategies in which it is desired to produce a gene product in one cell which is to be delivered to another location in the body. Secretion vectors may also facilitate the purification of desired proteins.

The present invention also relates to expression vectors capable of directing the expression of an inserted gene in a desired spatial or temporal manner or at a desired level. Such vectors may include sequences upstream of the EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids, such as promoters or upstream regulatory sequences.

The present invention also comprises fusion vectors for making chimeric polypeptides comprising a first polypeptide and a second polypeptide. Such vectors are useful for determining the cellular localization of the chimeric polypeptides or for isolating, purifying or enriching the chimeric polypeptides.

The EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids may also be used for gene therapy to control or treat genetic diseases. In the case of secreted proteins, signal peptides may be fused to heterologous proteins to direct their extracellular secretion.

Bacterial clones containing Bluescript plasmids having inserts containing the sequence of the non-aligned 5'ESTs, also referred to as singletons, and sequences of the 5'ESTs which were aligned to yield consensus contigated 5' ESTs are presently stored at 80°C in 4% (v/v) glycerol in the inventor's laboratories under internal designations. The non-aligned 5'ESTs are those which comprise a single EST from a single tissue in the listing of Table V. The inserts may be recovered from the stored materials by growing the appropriate clones on a suitable medium. The Bluescript DNA can then be isolated using plasmid isolation procedures familiar to those skilled in the art such as alkaline lysis minipreps or large scale alkaline lysis plasmid isolation procedures. If desired the plasmid DNA may be further enriched by centrifugation on a cesium chloride gradient, size exclusion chromatography, or anion exchange chromatography. The plasmid DNA obtained using these procedures may then be manipulated using standard cloning techniques familiar to those skilled in the art. Alternatively, a PCR

can be performed with primers designed at both ends of the inserted EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids. The PCR product which corresponds to the EST-related nucleic acids, fragments of EST-related nucleic acids, positional segments of EST-related nucleic acids, or fragments of positional segments of nucleic acids can then be manipulated using standard cloning techniques familiar to those skilled in the art.

One embodiment of the present invention is a purified nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and sequences complementary to the sequences of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622.

Another embodiment of the present invention is a purified nucleic acid comprising at least 10, 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, 50, 75, 100, 200, 300, 500, or 1000 consecutive nucleotides, to the extent that fragments of these lengths are consistent with the specific sequence, of a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and sequences complementary to the sequences of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622.

A further embodiment of the present invention is a purified nucleic acid comprising the coding sequence of a sequence selected from the group consisting of SEQ ID NOs. 24-811.

Yet another embodiment of the present invention is a purified nucleic acid comprising the full coding sequences of a sequence selected from the group consisting of SEQ ID NOs. 766-792 wherein the full coding sequence comprises the sequence encoding the signal peptide and the sequence encoding the mature protein.

Still another embodiment of the present invention is a purified nucleic acid comprising a contiguous span of a sequence selected from the group consisting of SEQ ID NOs. 766-792 which encodes the mature protein.

Another embodiment of the present invention is a purified nucleic acid comprising a contiguous span of a sequence selected from the group consisting of SEQ ID NOs. 24-728 and 766-792 which encodes the signal peptide.

Another embodiment of the present invention is a purified nucleic acid encoding a polypeptide comprising a sequence selected from the group consisting of the sequences of SEQ ID NOs. 812-1599.

Another embodiment of the present invention is a purified nucleic acid encoding a polypeptide comprising a sequence selected from the group consisting of the sequences of SEQ ID NOs. 1554-1580.

Another embodiment of the present invention is a purified nucleic acid encoding a polypeptide comprising a mature protein included in a sequence selected from the group consisting of the sequences of SEQ ID NOs. 1554-1580.

Another embodiment of the present invention is a purified nucleic acid encoding a polypeptide comprising a signal peptide included in a sequence selected from the group consisting of the sequences of SEQ ID NOs. 812-1516 and 1554-1580.

Another embodiment of the present invention is a purified nucleic acid at least 30, 35, 40, 50, 75, 100, 200, 300, 500 or 1000 nucleotides in length which hybridizes under stringent conditions to a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and sequences complementary to the sequences of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622.

Another embodiment of the present invention is a purified or isolated polypeptide comprising a sequence selected from the group consisting of the sequences of SEQ ID NOs. 812-1599.

Another embodiment of the present invention is a purified or isolated polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs. 1554-1580.

Another embodiment of the present invention is a purified or isolated polypeptide comprising a mature protein of a polypeptide selected from the group consisting of SEQ ID NOs. 1554-1580.

Another embodiment of the present invention is a purified or isolated polypeptide comprising a signal peptide of a sequence selected from the group consisting of the polypeptides of SEQ ID NOs. 812-1516 and 1554-1580.

Another embodiment of the present invention is a purified or isolated polypeptide comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, 50, 75, 100, 200, 300, 500, or 1000 consecutive amino acids, to the extent that fragments of these lengths are consistent with the specific sequence, of a sequence selected from the group consisting of the sequences of SEQ ID NOs. 812-1599.

Another embodiment of the present invention is a method of making a cDNA comprising the steps of contacting a collection of mRNA molecules from human cells with a primer comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive nucleotides of a sequence selected from the group consisting of the sequences complementary to SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, hybridizing said primer to an mRNA in said collection that encodes said protein reverse transcribing said hybridized primer to make a first cDNA strand from said mRNA, making a second cDNA strand complementary to said first cDNA strand and isolating the resulting cDNA encoding said protein comprising said first cDNA strand and said second cDNA strand.

Another embodiment of the present invention is a purified cDNA obtainable by the method of the preceding paragraph.

In one aspect of this embodiment, the cDNA encodes at least a portion of a human polypeptide.

Another embodiment of the present invention is a method of making a cDNA comprising the steps of obtaining a cDNA comprising a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, contacting said cDNA with a detectable probe comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive nucleotides of a sequence

selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and the sequences complementary to SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 under conditions which permit said probe to hybridize to said cDNA, identifying a cDNA which hybridizes to said detectable probe, and isolating said cDNA which hybridizes to said probe.

Another embodiment of the present invention is a purified cDNA obtainable by the method of the preceding paragraph.

In one aspect of this embodiment, the cDNA encodes at least a portion of a human polypeptide.

Another embodiment of the present invention is a method of making a cDNA comprising the steps of contacting a collection of mRNA molecules from human cells with a first primer capable of hybridizing to the polyA tail of said mRNA, hybridizing said first primer to said polyA tail, reverse transcribing said mRNA to make a first cDNA strand, making a second cDNA strand complementary to said first cDNA strand using at least one primer comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, and isolating the resulting cDNA comprising said first cDNA strand and said second cDNA strand.

Another embodiment of the present invention is a purified cDNA obtainable by the method of the preceding paragraph.

In one aspect of this embodiment, said cDNA encodes at least a portion of a human polypeptide.

In another aspect of the preceding method the second cDNA strand is made by contacting said first cDNA strand with a first pair of primers, said first pair of primers comprising a second primer comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and a third primer having a sequence therein which is included within the sequence of said first primer, performing a first polymerase chain reaction with said first pair of primers to generate a first PCR product, contacting said first PCR product with a second pair of primers, said second pair of primers comprising a fourth primer, said fourth primer comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive nucleotides of said sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, and a fifth primer, wherein said fourth and fifth hybridize to sequences within said first PCR product, and performing a second polymerase chain reaction, thereby generating a second PCR product.

One aspect of this embodiment is a purified cDNA obtainable by the method of the preceding paragraph.

In another aspect of this embodiment, said cDNA encodes at least a portion of a human polypeptide.

Alternatively, the second cDNA strand may be made by contacting said first cDNA strand with a second primer comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, hybridizing said second primer to said first strand cDNA, and extending said hybridized second primer to generate said second cDNA strand.

One aspect of the above embodiment is a purified cDNA obtainable by the method of the preceding paragraph.

In a further aspect of this embodiment said cDNA encodes at least a portion of a human polypeptide.

Another embodiment of the present invention is a method of making a polypeptide comprising the steps of obtaining a cDNA which encodes a polypeptide encoded by a nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs. 24-811 or a cDNA which encodes a polypeptide comprising at least 6, 8, 10, 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive amino acids of a polypeptide encoded by a sequence selected from the group consisting of SEQ ID NOs. 24-811, inserting said cDNA in an expression vector such that said cDNA is operably linked to a promoter, introducing said expression vector into a host cell whereby said host cell produces the protein encoded by said cDNA, and isolating said protein.

Another aspect of this embodiment is an isolated protein obtainable by the method of the preceding paragraph.

Another embodiment of the present invention is a method of obtaining a promoter DNA comprising the steps of obtaining genomic DNA located upstream of a nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and the sequences complementary to the sequences of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, screening said genomic DNA to identify a promoter capable of directing transcription initiation, and isolating said DNA comprising said identified promoter.

In one aspect of this embodiment, said obtaining step comprises walking from genomic DNA comprising a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and the sequences complementary to SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622. In another aspect of this embodiment, said screening step comprises inserting genomic DNA located upstream of a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and the sequences complementary to SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 into a promoter reporter vector. For example, said screening step may comprise identifying motifs in genomic DNA located upstream of a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and the sequences complementary to SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 which are transcription factor binding sites or transcription start sites.

Another embodiment of the present invention is a isolated promoter obtainable by the method of the paragraph above.

Another embodiment of the present invention is the inclusion of at least one sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, the sequences complementary to the sequences of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and fragments comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, 50, or 100 consecutive nucleotides of said sequence in an array of discrete ESTs or fragments thereof of at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, 50, or 100 nucleotides in length. In some aspects of this embodiment, the array includes at least two sequences selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, the sequences complementary to the sequences of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, and fragments comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, 50, or 100 consecutive nucleotides of said sequences. In another aspect of this embodiment., the array includes at least one, three, five, ten, fifteen, or twenty sequences selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, the sequences complementary to the sequences of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and fragments comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, 50, or 100 consecutive nucleotides of said sequences.

Another embodiment of the present invention is an enriched population of recombinant nucleic acids, said recombinant nucleic acids comprising an insert nucleic acid and a backbone nucleic acid, wherein at least 0.01%, 0.05%, 0.1%, 0.5%, 1%, 2%, 5%, 10%, or 20% of said insert nucleic acids in said population comprise a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and the sequences complementary to SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622.

Another embodiment of the present invention is a purified or isolated antibody capable of specifically binding to a polypeptide comprising a sequence selected from the group consisting of SEQ ID NOs. 812-1599.

Another embodiment of the present invention is a purified or isolated antibody capable of specifically binding to a polypeptide comprising at least 6, 8, 10, 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive amino acids of a sequence selected from the group consisting of SEQ ID NOs. 812-1599.

Yet, another embodiment of the present invention is an antibody composition capable of selectively binding to an epitope-containing fragment of a polypeptide comprising a contiguous span of at least 8, 10, 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 amino acids of any of SEQ ID NOs. 812-1599, wherein said antibody is polyclonal or monoclonal.

Another embodiment of the present invention is a computer readable medium having stored thereon a sequence selected from the group consisting of a nucleic acid code of SEQ ID NOs. 24-811 and 1600-1622 and a polypeptide code of SEQ ID NOs. 812-1599.

Another embodiment of the present invention is a computer system comprising a processor and a data storage device wherein said data storage device has stored thereon a sequence selected from the group consisting of a nucleic acid code of SEQID NOs. 24-811 and 1600-1622 and a polypeptide code of SEQ ID NOs. 812-1599. In one aspect of this embodiment the computer system further comprises a sequence comparer and a data storage device having reference sequences stored thereon. For example, the sequence comparer may comprise a computer program which indicates polymorphisms. In another aspect of this embodiment, the computer system further comprises an identifier which identifies features in said sequence.

Another embodiment of the present invention is a method for comparing a first sequence to a reference sequence wherein said first sequence is selected from the group consisting of a nucleic acid code of SEQID NOs. 24-811 and 1600-1622 and a polypeptide code of SEQ ID NOs. 812-1599 comprising the steps of reading said first sequence and said reference sequence through use of a computer program which compares sequences and determining differences between said first sequence and said reference sequence with said computer program. In some aspects of this embodiment, said step of determining differences between the first sequence and the reference sequence comprises identifying polymorphisms.

Another embodiment of the present invention is a method for identifying a feature in a sequence selected from the group consisting of a nucleic acid code of SEQID NOs. 24-811 and 1600-1622 and a polypeptide code of SEQ ID NOs. 812-1599 comprising the steps of reading said sequence through the use of a computer program which identifies features in sequences and identifying features in said sequence with said computer program.

Another embodiment of the present invention is a vector comprising a nucleic acid according to any one of the nucleic acids described above.

Another embodiment of the present invention is a host cell containing the above vector.

Another embodiment of the present invention is a method of making any of the nucleic acids described above comprising the steps of introducing said nucleic acid into a host cell such that said nucleic acid is present in multiple copies in each host cell and isolating said nucleic acid from said host cell.

Another embodiment of the present invention is a method of making a nucleic acid of any of the nucleic acids described above comprising the step of sequentially linking together the nucleotides in said nucleic acids.

Another embodiment of the present invention is a method of making any of the polypeptides described above wherein said polypeptides is 150 amino acids in length or less comprising the step of sequentially linking together the amino acids in said polypeptide.

Another embodiment of the present invention is a method of making any of the polypeptides described above wherein said polypeptides is 120 amino acids in length or less comprising the step of sequentially linking together the amino acids in said polypeptides.

Brief Description of the Drawings

Figure 1 is a summary of a procedure for obtaining cDNAs which have been selected to include the 5' ends of the mRNAs from which they derived. In the first step (1), the cap of intact mRNAs is oxidized to be chemically ligated to an oligonucleotide tag. In the second step (2), a reverse transcription is performed using random primers to generate a first cDNA strand. In the third step (3), mRNAs are eliminated and the second strand synthesis is carried out using a primer contained in the oligonucleotide tag.

Figure 2 is an analysis of the 43 amino terminal amino acids of all human SwissProt proteins to determine the frequency of false positives and false negatives using the techniques for signal peptide identification described herein.

Figure 3 summarizes a general method used to clone and sequence extended cDNAs containing sequences adjacent to 5'ESTs.

Figure 4 provides a schematic description of the promoters isolated and the way they are assembled with the corresponding 5' tags.

Figure 5 describes the transcription factor binding sites present in each of the promoters of Figure 4.

Figure 6 is a block diagram of an exemplary computer system.

Figure 7 is a flow diagram illustrating one embodiment of a process 200 for comparing a new nucleotide or protein sequence with a database of sequences in order to determine the homology levels between the new sequence and the sequences in the database.

Figure 8 is a flow diagram illustrating one embodiment of a process 250 in a computer for determining whether two sequences are homologous.

Figure 9 is a flow diagram illustrating one embodiment of an identifier process 300 for detecting the presence of a feature in a sequence.

Figure 10 is a table with all of the parameters that can be used for each step of extended cDNA analysis.

Detailed Description of the Preferred Embodiment

I. Obtaining 5'ESTs from cDNA libraries including the 5'Ends of their Corresponding mRNAs

The 5' ESTs of the present invention were obtained from cDNA libraries including cDNAs which include the 5'end of their corresponding mRNAs. The general method used to obtain such cDNA libraries is described in Examples 1 to 5.

EXAMPLE 1

Preparation of mRNA

Total human RNAs or polyA$^+$ RNAs derived from 29 different tissues were respectively purchased from LABIMO and CLONTECH and used to generate 44 cDNA libraries as described below.

C) Deletions in the sequence of a consensus contigated 5'EST to derive a preferred nucleic acid fragment are denoted by an "D", followed by a number indicating the first nucleotide position in a specific SEQ ID to be deleted in a string of deleted nucleotides or the position of the deleted nucleotide in the case of a single deleted nucleotide. Then there is a coma followed by number indicating the number of nucleotide(s) deleted from the sequence provided in the sequence ID. For example, SEQ ID NO: 5398; Position of preferred fragments: 56-780; Variant nucleotides D114,5 would indicate that a preferred polynucleotide fragment had the sequence of positions 56 to 780 of SEQ ID NO. 5398, except that the nucleotides in positions 114 to 118 had been deleted in the preferred polynucleotide as compared with the sequence of SEQ ID No. 5398.

The present invention encompasses isolated, purified, or recombinant nucleic acids which consist of, consist essentially of, or comprise a contiguous span of at least 8, 10, 12, 15, 18, 20, 25, 35, 40, 50, 70, 80, 100, 250, or 500 nucleotides in length, to the extent that a contiguous span of these lengths is consistent with the lengths of the particular polynucleotide, of a polynucleotide described in Table II, or a sequence complementary thereto, wherein said polynucleotide described in Table II is selected individually or in any combination from the polynucleotides described in Table II. The present invention also encompasses isolated, purified, or recombinant nucleic acids which consist of or consist essentially of a polynucleotide described in Table II, or a sequence complementary thereto, wherein said polynucleotide is selected individually or in any combination from the polynucleotides described in Table II. The present invention further encompasses isolated or purified polypeptides which consist of, consist essentially of, or comprise a contiguous span of at least 8, 10, 12, 15, 18, 20, 25, 35, 40, 50, 70, 80, or 100 amino acids encoded by a polynucleotide described in Table II.

Table II

SEQ ID NO.	Positions of Preferred Fragments	Variant nucleotides
35	1-423	S124, s; I135, a; S293, w; I363, a; S377, r; D424, 15
41	1-427	I117, m; S120, r; S124, g; D373, 1; S376, b; S378, b; I427, gggg; D428, 109
43	1-276	S114, m; S118, rg; S123, r; S139, nr; I142, t; D148, 1; D152, 1; I228, t; I276, gg; D277, 136
45	126-420	D1, 125; I420, ggg; D421, 100
46	1-255	S139, r; I145, r; S146, mm; S150, ar; S254, g; D256, 128
48	4-437	D1, 3; S49, a; S55, g; S79, a; S90, a; I437, tctctg
59	1-471	S26, a; S44, t; S48, t; S109, a; S191, t; S200, gc; S203, a; S210, g; S237, a; S240, g; S255, a; S272, a; S277, a; S279, a; S284, t; S297, g; S305, g; S316, a; I471, ggtca
66	1-428	I428, tactgggg

82	1-399	S251, t; S277, d; I399, aagccggg
84	5-488	D1, 4; S210, g; S293, a; S325, g; S339, a; S348, g; S353, g; S395, g; I488, cacca
93	1-508	I508, gattt
96	26-315	D1, 25; S28, a; S62, c; I315, cagatgg
97	4-460	D1, 3; S19, g; S31, g; S114, gt; S118, a; S123, tc; S127, c; S132, a; S186, g; S190, c; S203, t; S210, g; S232, c; I460, acgtt
105	1-281	S273, a; I281, g; D282, 211
114	10-315	I0, t; D1, 9; S91, m; S267, n; S276, w; S292, h; S295, m; I315, tggg; D316, 19
118	1-145	S57, d; S126, d; I145, ccctc
120	2-348	D1, 1; S104, t; I348, g; D349, 38
121	1-190	I121, c; I190, ccctt
123	1-353	I117, m; I186, w; S187, y; I353, caccgggg
124	1-249	I249, ggrvgggg
125	114-375	D1, 113; S206, wn; I231, a; I375, ccctagg
126	1-437	S297, cc; S307, tg; S312, a; S318, g; S341, a; S351, t; S353, g; S383, c; S387, a; D404, 1
136	82-428	D1, 81; I428, aaagtg
139	1-268	I268, gggaaggg
148	6-405	D1, 5; I405, ggtgt
159	1-230	S227, ta; I230, ccctggg
165	3-256	I0, tat; D1, 2; I17, c; S18, t; S111, d; I115, t; S123, r; I256, aaggcggg
170	1-280	I103, t; S104, c; I111, t; I280, cgttcggg
194	1-215	S50, s; S186, sn; S199, k; I215, gcagcggg
213	1-158	S128, m; I132, w; S143, d; I158, tgcccggg
223	3-431	D1, 2; S28, s; S79, c; S82, s; S308, nr; S328, nb; I431, ccggc
247	1-359	I76, gttt; I359, tccctgg
258	1-236	S72, r; S81, g; S197, s; I205, ss; S232, k; I236, acttcggg
264	5-283	D1, 4; S64, g; S122, m; S134, yy; I137, c; I151, t; I283, gttgc
269	1-143	S111, s; I143, ggggcggg
286	5-207	D1, 4; S204, a; S206, c; I207, gg; D208, 567
287	1-277	S114, r; I125, t; S131, ag; S256, tg; S259, tt; S262, at; S267, t; S269, c; S273, c; I277, ccggg; D278, 337
289	69-416	D1, 68; I416, agccaggg
289	1-278	S114, r; I125, t; S131, ag; S277, c; I278, cggg; D279, 138
292	20-254	D1, 19; I254, aaagagg
293	1-414	I414, tagcag
300	1-285	S16, m; S67, y; I285, baccacggg; D286, 1
349	23-431	D1, 22; I118, a; S214, y; I431, caactgg
350	3-386	D1, 2; S42, w; I263, c; I386, gggat
368	3-446	D1, 2; I446, tctct
385	1-193	I35, t; I108, t; I134, r; S135, a; S137, r; S143, w; I178, c; I193, gagcgggg
411	6-391	D1, 5; S17, r; S27, t; S334, y; D392, 244
412	1-185	S49, s; S127, s; I185, gctggg; D186, 150

415	2-229	D1, 1; S3, a; I229, caaatggg
435	1-386	S4, s; I386, ccggg
436	4-472	D1, 3; S61, sa; D238, 1; S239, s; I472, agtgtgg
437	1-340	I340, ggg; D341, 129
441	1-409	S109, smag; I409, cgcacggg
454	1-492	S72, nn; S115, t; S121, bwy; S181, yn; I492, gagtc
455	1-177	I14, w; I16, a; I177, gagctggg
459	1-311	S39, n; S74, rg; I311, accatggg
460	1-425	I425, agtac
461	5-420	D1, 4; I420, tcgtc
481	1-429	I10, w; S262, d; S333, n; I429, ctccaggg
489	1-414	D72, 1; S117, n; S396, d; I414, ggaca
496	1-215	I215, ttttcggg
501	1-430	S275, n; I430, aggat
502	91-413	D1, 90; I413, aaacgggg
504	21-420	D1, 20; S47, w; S83, n; I280, n; S281, na; S292, v; S314, sm; S368, ww; S373, w; I420, cccca
505	18-457	D1, 17; D36, 1; S182, g; S273, n; S283, a; S416, bh; I457, ctcga
514	1-303	I303, accca
515	1-455	S11, t; I12, n; S30, r; S256, wr; I333, t; I455, cataa
517	24-453	D1, 23; I453, agagcggg
519	1-275	I119, gt; S125, w; I129, w; S133, k; S137, k; S167, k; I275, gcccc
522	1-313	I313, agcgtggg
526	4-366	I0, t; D1, 3; I366, ggcccggg
530	1-434	S328, g; I434, aagat
535	1-379	S128, g; S162, m; D380, 5
561	2-341	D1, 1; I341, raagagg
568	1-246	I118, g; S137, g; I246, aaaccggg
570	1-207	I207, ttttt
576	1-288	I34, c; I288, cccgtgg
588	1-390	S218, a; S224, k; S314, dh; S358, s; D376, 1; I390, atg; D391, 23
597	31-274	D1, 30; S49, n; I274, tccatgg
606	1-354	I141, g; D174, 1; S229, rr; D355, 72
627	1-415	S7, a; I415, cattt
634	1-178	D179, 212
640	6-428	D1, 5; D429, 79
641	64-483	D1, 63; I165, d; D183, 1; S185, y; S253, t; D279, 2; S416, a; I483, atata
655	1-280	S58, c; I84, g; S88, k; S204, ac; S244, g; S247, g; I280, ggg; D281, 90
672	34-489	D1, 33; S316, k; S331, k; S333, w; S486, g; S488, c; D490, 4
687	116-473	D1, 115; S142, n; I473, cctcgggg
697	1-202	S142, s; S144, sr; S148, d; S152, d; I155, a; I164, a; S174, k; I202, gcc; D203, 291
708	8-384	D1, 7; S104, b; I384, gaaaa
710	1-167	S40, k; S49, db; I167, tatct

722	1-191	I125, c; I191, ttttt
723	1-316	I316, aggg; D317, 157
729	15-373	D1, 14; S139, t; I373, cgcag; D374, 99
730	29-372	D1, 28; I155, g; S192, ka; S333, d; I372, m; D373, 93
731	1-290	S10, kk; S30, b; S32, t; S92, t; S197, dy; S278, g; I290, aggg; D291, 55
732	8-277	D1, 7; I113, a; S127, w; I131, s; S132, r; S156, w; S160, r; S211, n; S215, w; I247, a; D278, 121
733	20-375	D1, 19; S306, sbs; I325, h; S326, nr; S338, ywd; S344, v; I375, aggg; D376, 68
734	1-359	D66, 1; D360, 14
735	25-322	D1, 24; S30, r; I193, a; I322, ccaaggg
736	9-181	D1, 8; S58, g; I181, aactaggg
737	1-160	S97, ta; I160, aggtc
738	1-227	D228, 7
739	45-514	D1, 44; S178, s; I182, c; S436, dmn; S461, v; S476, c; S506, t; D515, 75
740	11-388	D1, 10; I388, cgacaggg
741	1-478	S118, s; S125, a; I126, s; S134, k; S421, vn; I478, aatsc
742	217-553	I0, tt; D1, 216; S286, r; S294, m; S311, r; S317, s; S338, r; S442, dm; S469, h; S476, r; S485, s; S491, w; I495, ht; S496, v; S513, r; D521, 1; S536, m; D554, 199
743	1-459	I11, s; S258, m; I270, m; I304, c; I308, amta; S313, c; S438, v; I459, agggag
744	25-316	D1, 24; S315, g; D317, 95
745	21-283	D1, 20; I40, g; S41, c; D123, 1; S181, sr; S227, r; I283, ccgcg; D284, 121
746	1-256	D257, 173
747	1-179	S134, w; S138, w; S140, kt; I179, cacca
748	1-235	S46, t; I72, t; S189, cc; S222, c; D236, 148
749	2-370	D1, 1; S32, cg; D144, 1; S341, g; D371, 76
750	18-410	I0, aag; D1, 17; I410, aatcc
751	22-355	D1, 21; D148, 1; S150, c; S152, a; S313, n; D356, 181
752	1-139	S50, t; I118, g; I139, ccct
753	1-189	S26, r; S115, s; I121, r; S122, r; S128, s; S143, r; I146, w; S156, r; D190, 4
754	1-395	S212, wd; I395, cggca
755	19-460	D1, 18; S26, c; S156, a; S253, n; I460, tagaagg
756	2-142	D1, 1; I106, gc; S107, t; S110, c; I142, ccaccggg
757	28-296	D1, 27; I119, s; I122, t; S128, s; S255, t; S267, m; D297, 66
758	11-368	D1, 10; I200, g; S201, c; S281, d; S317, c; I368, ccatcggg
759	19-452	D1, 18; S421, w; I452, a
760	25-175	D1, 24; S34, yk; I175, ccggg; D176, 120
761	1-212	I212, cactcggg
762	1-374	S320, s; S349, a; D375, 249
763	8-152	D1, 7; I152, acggg; D153, 109

764	1-160	I127, g; I145, g; I160, cgcccggg
765	137-313	D1, 136; S272, m; I279, s; S310, t; I313, ggg; D314, 203
766	1-320	S278, ag; S281, cagacc; S288, ta; S291, caag; S296, c; S317, m; I320, cggg; D321, 306
767	6-336	I0, aa; D1, 5; S149, w; S245, y; D337, 137
768	1-374	S320, s; D375, 299
769	53-435	D1, 52; S59, b; S344, nnkw; D436, 104
770	24-448	D1, 23; S25, g; S411, w; S416, m; D449, 31
771	1-370	S3, c; S180, m; S275, r; D371, 122
772	1-388	I299, c; S326, c; D389, 8
773	1-143	S18, c; S66, a; I143, ggg; D144, 274
774	1-347	S194, a; S205, c; I347, ggg; D348, 107
775	5-207	D1, 4; S111, tg; S158, g; S171, c; S191, a; S204, a; S206, c; I207, gg; D208, 324
776	1-368	I200, c; S201, a; S291, ta; I332, c
777	5-207	D1, 4; S204, a; S206, c; I207, gg; D208, 262
778	39-342	D1, 38; S184, r; D343, 126
779	4-360	D1, 3; S13, m; S15, c; S22, s; S24, m; S48, r; S56, s; S335, c; S345, rs; I360, ggg; D361, 119
780	1-472	I347, c; D473, 32
781	116-426	D1, 115; S219, m; S424, g; D427, 118
782	1-391	S386, k; D392, 64
783	1-453	D109, 1; S110, y; S125, y; I128, g; S132, k; I453, ctctc
784	29-494	D1, 28; S72, r; D495, 93
785	99-461	D1, 98; S218, r; I461, gaccgggg
786	2-465	D1, 1; S8, y; S388, s; I398, g; S400, t; S403, at; S417, g; D466, 24
787	28-271	D1, 27; S99, t; S230, c; S266, ga; S269, c; I271, g; D272, 126
788	1-285	D280, 1; I285, g; D286, 310
789	1-209	S205, c; D210, 150
790	51-297	D1, 50; I297, ggggg; D298, 539
791	113-327	D1, 112; S218, g; I226, g; D280, 1; I327, cgcaggg; D328, 224
792	17-218	D1, 16; S58, t; S217, t; I218, gggg; D219, 219
793	11-92	D1, 10; S91, c; I92, a; D93, 258
794	9-431	D1, 8; I431, taagt
795	30-341	D1, 29; I341, a; D342, 175
796	1-442	S17, w; S19, wr; D35, 1; S134, t; S264, n; S322, nr; S369, s; S420, s; S422, y; I442, tcctcggg
797	1-420	S136, c; S150, c; I245, ccc; I420, ggagtg
798	25-316	D1, 24; S315, g; D317, 97
799	1-344	D345, 57
800	7-465	D1, 6; S59, k; S146, a; S186, krn; I465, gttca
801	121-422	D1, 120; I269, c; S419, cc; I422, gg; D423, 207
802	46-477	D1, 45; S132, bn; I477, actac
803	15-467	D1, 14; S45, k; S65, t; S418, ys; D452, 1; D468, 119
804	1-341	S42, t; S97, d; S326, gtg; S331, tgt; S336, a;

		S338, c; I341, cccccggg; D342, 218
805	2-409	D1, 1; S334, d; I409, aggg; D410, 161
806	5-384	D1, 4; I384, actaa
807	1-301	S113, a; S117, c; S123, t; D128, 1; D134, 1; S282, g; S284, a; I301, gacggagggg; D302, 70
808	2-314	D1, 1; S306, g; I314, ggg; D315, 121
809	1-394	S53, g; S228, n; S272, vk; I301, g; I358, m; S368, nb; S375, w; I383, mm; I388, yt; I394, nhaccggg
810	6-205	I0, a; D1, 5; I141, t; I205, ggg; D206, 630
811	6-270	D1, 5; I270, gggg; D271, 115
1600	1-247	S45, m; S114, k; I122, m; S123, yc; S158, rr; S221, k; I247, ccccaggg
1601	1-225	S109, bm; S195, m; I225, tgcacggg
1602	23-245	D1, 22; D138, 1; S139, s; S242, t; S244, g; I245, g; D246, 13
1603	1-303	S71, c; D277, 1; I303, ggagggg; D304, 38
1604	1-242	S47, w; S50, c; S81, h; S85, d; S91, k; S106, r; I242, tgtggg; D243, 50
1605	2-225	D1, 1; S20, k; S91, c; I225, ggg; D226, 132
1606	15-293	D1, 14; S156, g; S193, g; I200, t; I293, acaaaggg
1607	1-361	S323, c; I361, cccca
1608	1-151	I151, taagggg; D152, 154
1609	1-242	S55, s; I135, a; S152, h; I242, cagtaggg
1610	1-196	I151, w; S190, k; I196, cctgtgg
1611	1-228	S115, k; S174, rk; I228, cgtttggg
1612	1-221	S108, v; I221, tgatcggg
1613	1-281	I66, w; I137, a; D282, 79
1614	1-171	S53, k; S76, k; I80, k; S81, kw; S86, r; S92, k; S126, k; I171, gccgagg
1615	2-193	D1, 1; S67, c; I121, s; S122, mm; S126, g; S130, r; S146, r; S156, gm; I193, cctca
1616	1-349	S251, ww; S259, rs; S275, k; I279, w; S285, y; S292, y; I320, m; I331, m; I338, w; I341, s; I349, accccggg
1617	1-129	I118, t; D130, 26
1618	1-184	D9, 1; D185, 1
1619	1-169	I122, t; I169, gcccaggg
1620	1-187	S106, k; S118, m; S122, cg; S132, k; D188, 59
1621	1-153	D125, 1; I131, ttt; S152, t; I153, gg; D154, 127
1622	1-400	S43, s; I126, g; I129, y; S353, d; I400, tatat

EXAMPLE 16

Categorization of 5' ESTs and Consensus Contigated 5'ESTs

The nucleic acid sequences of the present invention (SEQ ID NOs. 24-811 and 1600-1622) were grouped based on their homology to known sequences as follows. All sequences were compared to EMBL release 57 and daily releases available at the time of filing using BLASTN. All matches with a minimum of 25 nucleotides with 90% homology were retrieved and used to compute Tables IV and V.

In some embodiments, 5'ESTs or consensus contigated 5'ESTs nucleic acid sequence do not match any known vertebrate sequence nor any publicly available EST sequence, thus being completely new.

In other embodiments, 5'ESTs or consensus contigated 5'ESTs match a known sequence.

Tables III and IV gives for each sequence of the invention in this category referred to by its sequence identification number in the first column, the positions of their preferred fragments in the second column entitled "Positions of preferred fragments." As used herein the term "polynucleotide described in Table III" refers to the all of the preferred polynucleotide fragments defined in Table III in this manner, and the term "polynucleotide described in Table IV" refers to the all of the preferred polynucleotides fragments defined in Table IV in this manner. The present invention encompasses isolated, purified, or recombinant nucleic acids which consist of, consist essentially of, or comprise a contiguous span of at least 8, 10, 12, 15, 18, 20, 25, 35, 40, 50, 70, 80, 100, 250, or 500 nucleotides in length, to the extent that a contiguous span of these lengths is consistent with the lengths of the particular polynucleotide, of a polynucleotide described in Table III or Table IV, or a sequence complementary thereto, wherein said polynucleotide described in Table III or Talbe IV is selected individually or in any combination from the polynucleotides described in Table III or Talbe IV. The present invention also encompasses isolated, purified, or recombinant nucleic acids which consist of or consist essentially of a polynucleotide described in Table III or Table IV, or a sequence complementary thereto, wherein said polynucleotide is selected individually or in any combination from the polynucleotides described in Table III or Table IV.

Table III

SEQ ID NO	Positions of preferred fragments
24	1-251
25	1-83
28	227-276
29	1-27
30	130-242, 283-315, 365-461
32	314-399
33	89-321
34	1-38
35	1-52, 171-222
36	1-30, 408-441
37	1-138
39	115-140
40	1-97
41	1-112
42	1-177
46	1-38
48	376-400
51	400-466
54	1-259
55	189-320

56	265-457
58	246-469
59	81-123, 418-444
60	1-348
61	78-123, 418-457
62	386-439
63	1-214
64	109-297
65	1-370
66	92-428
68	1-180
69	165-259
70	1-178
71	1-27
72	1-179
73	1-65, 107-192
75	1-314
77	263-388
78	1-64
79	1-149
80	101-142, 302-380
82	1-192
83	1-398
85	1-290
86	1-118, 149-336
87	1-262
88	1-149
89	1-315
90	1-74
91	1-335, 364-423
92	1-316
93	338-508
94	179-321
95	219-402
96	26-315
97	348-460
98	1-230
99	391-467
101	214-336
102	1-289
103	1-383
104	1-211
105	1-36
106	1-126
107	1-49
108	294-336
109	1-128
111	1-154
112	407-441
113	1-80, 139-184
114	10-79
116	1-292
117	1-304

119	1-288
120	2-348
121	1-122
123	188-353
124	1-249
125	295-375
128	1-244
129	1-232
130	196-312
131	178-276
132	37-174
133	1-344
134	1-244
135	1-217
136	82-428
137	1-29, 103-155, 274-434
138	1-395
139	1-268
140	1-170
141	1-396
142	1-73, 227-357
143	1-159
144	1-433
145	61-116
146	1-71, 179-205
147	177-300
149	1-146
151	1-166
152	1-382
153	1-208
154	121-251
155	1-147
157	1-115
158	1-175
159	1-44, 80-230
160	1-346
161~	1-277
162	1-235
163	1-34
164	1-195
165	19-78, 175-217
166	1-209
167	1-65
168	128-218
169	49-245
170	179-280
171	1-103
172	1-218
173	1-380
174	1-139
175	1-122
176	1-300
177	1-466

179	1-86
180	1-245
181	1-241
182	1-263
183	1-170
184	58-106, 399-443
185	1-427
186	1-365
187	1-260
188	1-172
189	1-150
190	161-271, 301-339
191	1-91
192	1-264
193	1-246
194	1-150
195	1-209
196	1-363
197	1-155
198	1-135
200	1-125
201	1-210
202	1-338
203	1-188
204	228-347
205	1-440
206	56-221
208	1-422
209	169-195
210	1-363
211	1-368
212	1-448
213	1-134
214	1-193
215	1-214
216	1-134
218	1-189
219	1-248
220	1-115
221	1-113
222	1-370
224	1-251
225	1-198
226	45-141
227	1-206
228	1-480
229	1-144
230	1-42, 281-351, 432-457
231	1-112
233	1-301
234	1-109
235	1-393
236	1-222

237	1-154
238	1-439
239	112-137
240	1-194
241	1-44
242	1-242
244	1-324
245	1-38, 217-280
246	1-60
247	77-359
248	1-236
249	1-342
250	80-382
251	1-303
252	62-259
253	1-165
254	1-328
255	1-320
256	1-305
257	1-181
258	116-174
259	1-265
260	1-272
261	1-62
263	1-371
266	1-274
267	1-342
268	364-427
269	31-143
270	1-79
271	1-121
272	229-292
273	1-158
274	1-113
275	1-254
276	1-333
277 ~	1-130
278	1-184
279	1-265
280	1-188
281	1-177
282	1-336
283	1-294
284	1-171
285	1-297
288	1-42
290	1-170
292	20-155
294	1-334
295	1-375
296	1-226
297	1-232
299	40-139

300	1-285
301	1-242
302	1-136
303	1-175
304	1-493
305	1-214
306	89-458
307	1-328
308	1-380
309	1-236
310	1-357
311	1-470
312	1-187
313	1-159
315	1-162
316	1-404
317	1-450
318	1-395
319	1-257
320	56-325
321	1-201
322	1-159
323	1-420
324	1-210
325	1-192
326	88-181
327	1-185
328	128-210
330	1-223
331	1-362
332	1-89
334	1-188
335	1-115
336	1-300
337	1-307
338	1-123
339~	1-297
340	1-34
341	1-44
342	1-37
343	141-169
344	1-112
345	1-235, 266-349
346	1-191
347	1-229
348	1-210
350	139-266
351	1-307
352	1-170
353	1-293
354	30-161, 192-331
355	1-93
356	1-178

357	1-107
358	1-29, 168-209
359	1-298
360	1-193
362	1-360
363	1-45, 100-212
364	39-170, 202-242
365	1-248
366	1-351
367	1-208
368	228-446
369	1-62
370	1-132
371	1-127
372	1-196
373	1-148
374	1-126
375	1-112
376	1-146
378	1-143
379	1-261
380	202-228
382	1-151
383	1-45
384	1-190, 250-456
385	1-55, 141-181
386	1-281
387	1-111
388	1-374
389	1-192
390	1-371
392	1-303
394	1-126
395	1-329
396	1-99
397	1-316
398	1-251
399	1-120
401	1-206
402	1-330
403	1-311
405	1-153
406	1-206
407	1-479
408	1-289
410	229-321
413	1-158
415	95-229
416	1-265
417	1-228
418	1-225
419	207-293
420	1-194

421	1-90
422	1-161
423	1-420
424	1-432
425	1-276, 309-419
426	1-232
427	1-81
428	1-96
429	1-165
431	1-58, 186-237, 327-354
433	1-65
434	1-83
435	1-386
436	405-447
438	1-106
439	45-105, 168-255, 284-447
441	1-409
442	1-320
443	1-256
444	1-284
445	1-240
446	1-149
447	1-360
448	1-123
449	1-94
450	1-302
452	1-349
453	1-270
454	1-492
455	17-105
456	1-102
457	1-108
458	1-285
459	1-311
460	1-191
461	312-420
462~	1-257
463	1-117
464	1-142
466	1-235
467	1-29
468	1-41
469	1-438
470	1-131
471	1-211
472	1-150
473	1-352
474	1-141
476	1-232
478	1-201
479	1-151
480	1-104
481	7-429

482	1-385
486	1-226
488	1-296
489	1-72, 323-377
491	1-348
492	33-126
493	1-300
494	1-295
495	1-244
496	1-215
497	1-255
499	1-174, 384-474
500	1-50, 102-241
501	153-430
502	91-132
503	1-64
504	21-63, 356-420
505	37-68, 187-234
506	1-315
507	101-208
510	1-402
511	1-343
512	1-140, 170-246, 276-420
513	1-324
514	1-303
515	13-340
516	1-263, 293-360
518	1-245
519	111-275
520	62-182
521	1-218
523	1-502
524	1-118
525	1-276
526	223-366
527	1-428
528	297-342
529	1-244
530	1-88, 375-434
531	1-406
533	1-149
534	1-145
535	1-116
536	1-207
537	1-394
538	1-415
539	1-160
540	1-327
541	1-38, 73-396
542	1-247
543	1-221
544	1-375
545	1-376

546	1-109
547	1-160, 223-306
548	1-148
551	1-231
552	1-229
553	1-232
554	1-141
555	1-376
556	1-279
557	1-340
558	1-51
559	1-354
562	1-188
563	1-229
564	184-352
566	308-341
567	1-218
568	1-79
569	1-142
570	1-207
571	1-373
572	1-195
573	1-352
574	1-121
575	1-222
576	151-288
577	1-264
578	1-205
580	1-171, 273-328
581	1-356
582	1-239
583	1-144
584	1-282
585	1-338
586	1-436
588	1-380
589	1-60
590	1-178
592	1-66
593	1-215
594	1-161
596	1-407
597	31-83
598	1-417
599	1-329
600	1-311
601	1-61, 99-214
602	1-154, 197-463
603	135-269
604	1-351
605	1-195
608	1-357
609	1-201

612	1-176
613	1-342
615	1-272
616	1-114
617	1-46
618	1-208
619	1-257
620	1-28
621	1-26
622	1-221
623	1-432
624	1-233
625	1-26
627	1-43
628	1-318
629	1-170
630	1-196
631	248-339
632	1-433
633	1-154
634	1-41
635	1-137
636	1-172
637	1-253
638	1-185
639	1-206
641	334-483
642	1-309
643	1-75, 162-213
644	107-211
645	1-98
646	1-347
647	1-49, 81-143
648	1-232
649	74-133
650	1-37
651	1-276
652	1-170
653	1-178
654	1-121
656	1-197
657	1-246
659	1-197
660	116-172
661	1-411
662	1-146
663	1-65
664	1-182
665	1-320
666	1-273
667	1-149
668	1-122
670	1-160

671	1-137
673	1-263
674	1-263
675	1-107
677	1-441
678	134-191
679	1-235
680	1-26
682	1-58, 269-328
683	1-447
684	1-217
685	1-132
686	1-60
688	1-107
689	132-221, 327-377
690	1-388
691	1-141, 171-408
692	1-322
693	1-153
695	1-455
698	1-58, 117-174
699	240-300
700	1-159
701	1-69
702	1-175
703	1-298
704	1-136
705	1-168
706	1-419
707	1-382
708	8-245, 296-384
709	1-149
710	1-167
711	1-35
712	1-80, 116-156, 206-241
713	33-376
714	1-304
715	1-242
717	1-145
718	1-350
720	1-257
721	1-360
722	1-191
724	1-139
726	1-207
727	99-164
728	1-321
730	156-372
731	1-109, 256-290
735	25-192
737	1-160
738	1-227
739	441-514

742	217-280
743	10-275
747	1-179
749	2-31, 139-168
750	349-410
752	1-119
753	1-121
754	1-28
760	25-175
761	1-212
763	8-75
766	1-59, 102-248, 295-320
769	53-85
771	1-370
774	1-347
776	1-200
778	39-342
779	4-28
780	1-49, 407-472
781	116-426
782	1-59
783	1-53, 219-453
784	29-53, 219-263, 426-494
785	99-347, 386-461
786	2-28
788	1-279
789	1-58
790	226-268
792	129-218
794	265-431
796	5-86
797	1-34
799	1-344
802	46-477
806	64-384
807	135-301
808	2-314
810	6-39
1600	1-25
1601	1-225
1602	23-139
1603	1-294
1606	15-44
1607	1-361
1611	85-228
1612	1-221
1613	138-281
1614	65-171
1615	2-142
1616	1-46
1617	1-95
1620	1-187
1621	1-136

1622	32-280, 311-400

5

Table IV

SEQ ID NO	Positions of Preferred Fragments
35	1-52
41	1-115
45	1-47
46	1-33
66	400-428
82	83-149
93	399-508
105	1-36
114	1-79
120	1-386
121	1-190
124	1-249
125	295-328
139	1-81, 125-268
159	1-139, 180-230
165	1-78
170	179-205, 248-280
194	1-150
213	1-158
247	1-104, 155-183, 280-359
269	31-143
350	139-386
368	228-446
385	1-72, 143-193
415	95-229
435	1-386
436	446-472
441	1-361
454	1-349
455	1-105
459	35-161, 200-311
460	1-26, 56-140
481	1-429
489	1-84
496	1-44, 84-215
501	153-430
502	1-91
504	1-63

505	1-68
514	1-303
515	237-351
519	1-145
526	231-366
530	1-88
535	1-55
570	76-207
576	168-218, 261-288
588	1-331
597	1-83
627	1-43
634	1-41
641	1-55, 334-483
672	1-34
687	1-129
708	1-245, 296-384
710	1-26, 104-167
722	1-191
730	1-465
731	1-43
735	1-91
737	1-160
738	1-186
739	1-48
742	1-62, 99-248
743	1-315, 412-459
744	1-31
747	1-63
749	1-32
750	1-38
752	1-139
753	1-193
754	1-28
759	1-38
760	1-115
763	1-62
765	1-126
769	1-85
770	1-40
771	1-148
774	1-134
775	265-531
776	71-203
777	333-469
778	144-468
779	1-28
780	1-49
781	1-102
782	1-59
783	1-53
784	1-220, 262-390
785	1-339, 408-461

786	1-28
789	1-58
791	1-126
792	1-31, 129-220
793	1-31
794	355-431
795	1-33
797	1-31
798	1-31
799	1-401
801	1-117
802	1-92
806	64-384
807	1-331
808	1-351
810	1-39
1600	1-25
1603	1-341
1606	1-31
1607	1-361
1608	164-305
1611	85-228
1612	1-221
1613	112-360
1614	1-171
1615	94-193
1617	1-155
1620	1-246

III. Evaluation of Spatial and Temporal Expression of mRNAs Corresponding to the 5'ESTs, Consensus Contigated 5'ESTs, or EST-related nucleic acids

EXAMPLE 17

Expression Patterns of mRNAs From Which the 5'ESTs were obtained

Each of the SEQ ID NOs. 24-811 and 1600-1622 was also categorized based on the tissue from which its corresponding mRNA was obtained, as follows.

Table V shows the spatial distribution of each nucleic acid sequence of the invention (SEQ ID NOs. 24-811 and 1600-1622) referred to by its sequence identification number in the first column. In the second column entitled tissue distribution, the spatial distribution is represented by the number of individual 5'ESTs used to assemble the consensus contigated 5'ESTs for a given tissue. Each type of tissue listed in Table V is encoded by a letter. The correspondence between the letter code and the tissue type is given in Table VI.